Academy of Romanian Scientists
Journal of Knowledge Dynamics

Some Logical Aspects of the Concept of Artificial Intelligence

Emil DINGA[1]
[1] Centre for Financial and Monetary Research „Victor Slăvescu", Romanian Academy House, Building West, Floor 5, Room 5715, Romania, ORCID 0000-0002-3750-3248, emildinga2004@gmail.com (corresponding author)

Received: October 11, 2024
Revised: October 23, 2024
Accepted: October 31, 2024
Published: December 16, 2024

Abstract: The paper proposes a logical definition of the concept of artificial intelligence (AI), based on sufficiency predicates, by moving from genus to species. The nature of the intra-contingent need for AI, respectively the logical characteristics of AI, is shown. The digital/analogical relationship is discussed and, on this basis, issues in the AI "zone" such as: conscious/subconscious, free will, self-learning are examined. Finally, the issue of human protection against AI is assessed, respectively that of AI protection against humans and AI itself.

Keywords: predicate; automaton; logically living; consciousness; self-learning; free will.

The Introduction

Artificial intelligence (AI) is an automaton, i.e., an artifactual entity, which is capable of lasting in time and carrying out, during this duration, operations assimilable to a logical life (as opposed to biological life). To give a logical definition to IA we will have to find its sufficiency predicates. We will discuss two concepts related to AI: (a) the automaton; (b) the living in the logical sense (or the logically living).

Basic concepts

Automaton (A)

Of the three possible entities in the Universe – things, properties, relations – (Dinga, 2024), the automaton is a thing. As a thing, it has, of course, a series of properties, i.e., it generates a series of relations. We will refer to these, in the following, as its sufficiency predicates, i.e., the minimum number of attributes/characteristics that, through cumulative verification, qualifies the entity in question as an automaton:

(i) (A_1) is *artifact*: this means that it is a construct, either physical or ideal (logical, mathematical), and not a product of nature;

(ii) (A_2) is (or contains) a *self-controllable operational program* (*Nota bene*: even if, of course, being an artifact, the program has an external origin – we will refer to the so-called self-learning phenomenon below);

(iii) (A_3) is *capable of self-replication*: this capability is included in the program (*Nota bene*: this distinct sufficiency predicate is needed because its presence in the immediately preceding sufficiency predicate is not implied).

So, strictly logically, the automaton (A) is generated (comes into existence) by the logical conjunction of its sufficiency predicates:

$$A = (A_1) \wedge (A_2) \wedge (A_3) = \wedge_{i=1}^{3} A_i$$

How to cite
Dinga, E. (2024). Some Logical Aspects of the Concept of Artificial Intelligence. *Journal of Knowledge Dynamics,* Vol. 1, No. 2, p21-32. https://doi.org/10.56082/jkd.2024.2.21 ISSN ONLINE 3061-2640

Logically living (LL)

Analogously to the way in which we have proceeded with the definition of automaton, we will also proceed with the definition of the living logical system, or, equivalently, of life in the logical sense (Dinga, 2020). The sufficiency predicates of logically living (LL), or of a logical living system are the following:

(i) (L_1) is an *automaton*: it verifies the three sufficiency predicates (A_i), previously described;

(ii) (L_2): is an entity capable of *reaction*, either in response to internal actions or in response to external (trans-membranous) actions. We note that by reaction must be understood a very wide spectrum of inter-actions, either between the component elements of the entity in question, or with the environment of that entity:

- *ex post* evaluative reactions: the ability to examine, measure, characterize, predict (to different extents) past actions, to which a response action (a reaction) can oppose – these evaluative reactions may (or may not) be of actional kind;
- *ex-ante* evaluative reactions: the ability to examine, measure, characterize, predict (to different extents) future actions, to which a response action (a reaction) can oppose – these evaluative reactions may (or may not) be of actional kind;
- *classificatory* reactions: reactions consisting in making typologies/nomenclatures of the entities in the environment of the entity in question, so that a "map" is built for the environment from the perspective of interest criteria of the concerned entity (primarily, of survival);
- *definitional* reactions: reactions consisting of conceptualizations (regardless of the degree of cognitive sophistication) regarding the entities in the environment of the entity in question – particular forms of conceptualizations are: naming, establishing predicates of sufficiency (or of necessity), defining, interpreting of various natures and various degrees of cognition, etc.

(iii) (L_3) is an *evolutionary* entity: an evolutionary entity is an entity that undergoes changes in the direction of the fitness required by its environment (*Nota bene*: rigorously, any evolution is a co-evolution), i.e., an entity capable of designing and carrying out reactions which have "arrow", i.e., the fitness's chreode itself.

Then, it can be written:

$$LL = (A) \wedge (L_2) \wedge (L_3) = (A_1) \wedge (A_2) \wedge (A_3) \wedge (L_2) \wedge (L_3)$$

<u>Additional comment</u>: biologically living (BL) is not equivalent to logically living (LL), because the former violates the predicate A_1 (it is not an artifact), more precisely, biologically living is the conjunction of the following sufficiency predicates:

$$BL = (A_2) \wedge (A_3) \wedge (L_2) \wedge (L_3)$$

Artificial intelligence (AI)

We consider artificial intelligence to be a *non-axiological* logically living. In other words, a logically living that is forbidden to be value-permissive, i.e., that is axiologically opaque, can function as an artificial intelligence. We will introduce the following sufficiency predicates regarding the definition of AI:

(i) (AI_1) is an *opaque* (non-permissive) entity to *values*, including ethical values: the entity in question is not capable of autonomous axiological judgments (of course, by sophisticated programs, judgments can be formulated but they are para-axiological. The criterion to identify a genuine-axiological judgment is the following: a judgment is para-axiological if it is deductively predictable. *Nota bene*: the question may be asked: what can be said about an inductively predictable judgment? Our answer is: induction cannot produce predictions, but only forecasts – the distinction between prediction and forecast is: (a) a prediction is a statement about the future formulated as a lemma of a hypothesis/theory; (b) a forecast is a statement about the future formulated as a phenomenological extrapolation of (or by analogy with) the past;

(ii) (AI_2): is an entity that *encodes* communication signs (of any nature and level), both within the entity in question, and in the relationships between the entity and its environment, in a *digital* system (Kreutzer *et al.*, 2024). Encoding in the digital system is an encoding that presents the following essential characteristics:
   - it is a *discrete coding* system – until the current technological moment, this coding uses the binary system (with two signals: "0" and "1" or, "absence" and "presence"), physically instrumentalized by the electric current;
   - as a result of the discrete coding, it is a *mono-semantic* system – any sign (or combination of signs: word, sentence, judgment, reasoning) has a precise (exact) meaning, i.e., a unique, non-problematic referential/denoted. Consequently, digital signs are devoid of connotation/signification. *Nota bene*: in essence, connotation/signification is the idiosyncratic hypostasis of denotation/meaning.

From a logical point of view, AI is a conjunction between the logically living and the two new sufficiency predicates::

$$AI = (LL) \wedge (AI_1) \wedge (AI_2), \text{ or}$$

$$AI = (A_1) \wedge (A_2) \wedge (A_3) \wedge (L_2) \wedge (L_3) \wedge (AI_1) \wedge (AI_2)$$

Additional comment: we can now define natural intelligence (NI): it is the logically living that violates the sufficiency predicates AI_1 and AI_2:

$$NI = (A_1) \wedge (A_2) \wedge (A_3) \wedge (L_2) \wedge (L_3) \wedge \{[(AI_1) \wedge (\overline{AI_2})] \vee [(\overline{AI_1}) \wedge (AI_{2_2})] \vee [(\overline{AI_1}) \wedge (\overline{AI_2})]\},$$

where:

- $\overline{AI_1}$: the entity is transparent/permissive with respect to values or axiology;
- $\overline{AI_2}$: the entity encodes the signs in *analogical* system. Analogical coding of signs implies, through the logical negation of discrete system coding, the following:
   - the coding is *continuous* – the signs cannot be distinguished from each other in a crispy way (i.e., "without rest", or in a non-ambiguous way), as happens in the case of discrete/digital coding), but with a shadow of semantic overlap, imprecisely, problematically or even ambiguously);
   - as a result of continuous coding, it is a *poly-semantic* system or, more... "exactly", fuzzy-semantic, i.e., any sign (or combination of signs: word, sentence, judgment, reasoning) has a meaning, i.e., a referential /denoted non-unique, so, problematic. Accordingly, analogical signs contain, to the highest degree, connotation/signification.

The nature of AI – an intra-contingent necessity

Human society is not an artificial intelligence species

Human society, as a whole, is an artifact (a macro-artifact, more precisely). In fact, as the great theories of the origin of the state show, human society is based on the social contract. The foundation of society (and, respectively, the state) on the social contract which, subsequently, formalizes its principles in the Constitution (as a law of positive laws) gives society the character of a macro-artifact. It is obvious that the existence of the Constitution represents, in fact, a program of organization, operation and self-control of society, ensuring, at the same time, its self-replication. So, at the macro level, within Nature (as a general environment), human society (respectively the state) constitutes an *automaton*. The ability of society to react to its environment is implicit, and the evolutionary character (respectively, co-evolutionary in relation to the non-anthropic natural environment) of society no longer needs additional arguments. In conclusion, human society can be considered, from a logical point of view, as a *logically living*.

However, human society is not only not opaque to values and axiology, but is the only value-producing entity – human society is always (and primarily) an axiological society (*Nota bene*: one would say it is a symbolic society, but I think that qualification is a little bit careless). So, although society verifies the sufficiency predicates of a logically living, it violates one of the sufficiency predicates of artificial intelligence, so human society cannot, as such, be considered a species of artificial intelligence. On the other hand, according to the above, human society cannot be considered a species of natural intelligence, because, at the societal level, the sufficiency predicate $\overline{AI}_2$ is not verified as such – in the societal framework we also have systems of artificial intelligence, e.g., robots, automated systems of various kinds or technological levels, and natural intelligence systems, e.g., human beings (*Nota bene*: we do not yet know whether other biologically living systems, apart from human beings, verify the predicates of natural intelligence). Human society can be said to be a system of *mixed intelligence* or...*artiral intelligence*– from artificial and natural.

Intra-contingent necessity and artificial intelligence

(a) conceptual generalities

From a modal logic point of view, events occur either necessarily or contingently. An event is necessary if it is impossible for it not to occur (so, necessity includes possibility), and an event is contingent if it is possible but not necessary to occur. If we note: N – necessity, P – possibility, R – realization, C – contingency, then we have:

$$C(R) \leftrightarrow [P(R)] \wedge [\overline{N}(R)]$$

A dynamic relationship occurs between necessity and contingency, so that reality (either objective, subjective, or objectified – i.e., Popper's three worlds) presents itself as a *sui generis* combination of necessity and contingency. Contingency sub-episodes can occur within necessity episodes (e.g., suicide, which is contingent related to death, the latter being necessary) – and we call it intra-necessity contingency, and, symmetrically, this time within contingency episodes, necessity sub-episodes may occur (e.g., boiling water for tea is a contingent event but, after the temperature of the water reaches 100^0C, the water necessarily boils) – and we call it intra-contingency necessity.

(b) artificial intelligence as an intra-contingency necessity

Human being, as a biological entity (i.e., biologically living) is a contingency – the appearance of human being, through natural evolution, is not a necessity, it could appear or not appear on planet Earth, so the event of the appearance of human being is a contingent fact – possible and non-necessary. The traits (sufficiency predicates) of human being, as a species – among them: intelligence and, above all, consciousness – led him/her to develop exo-somatic tools that, qualitatively and quantitatively, entered a loop of positive feedback logic (self-catalysis). In other words, human being – a contingently initiated entity – necessarily developed an exo-somatic civilization or a technological one *(Nota* bene: civilization, like culture, represents a macro-artifact). One of the stages of this self-catalyzing process is exactly the artificial intelligence. So, from a logical perspective, artificial intelligence is an intra-contingency necessity.

The fundamental characteristics of artificial intelligence

Artificial intelligence, as defined and described above, involves a number of fundamental traits or characteristics. We consider the following ten characteristics of artificial intelligence, as resulting from its previous logical definition, to be decisive and relevant:

- is *algorithmic*: based on the predicates of automaton, AI is exclusively algorithmic, i.e., purely deterministic; the antonym of algorithmicity is randomness or intuitiveness, so AI is *non-intuitive*;
- is *sequential*: AI processuality, of whatever type, is a sequential or linear processuality (loops, which are common in AI, for example, sub-programs that are repeated based on certain logical conditions, does not introduce non-linearity at all); by contrast, natural intelligence is non-linear, non-sequential. *Nota bene*: the technical arrangements by which several AI entities are connected to each other and operate, thus, in parallel, do not at all mean that they operate non-sequentially (non-linearly) – the parallelism of their operation refers only to the level of performance (for example, computing speed), not to the way of that performance that remains sequential, being a case of *space-sharing* – analogously to the so-called *time-sharing*;

- is *non-hermeneutic*: this means that the AI is not capable of interpretations, but only of mono-semantic logical deductions, perfectly predictable by means of the operating programs of the entities in question, that is, of what is called reaction; non-hermeneuticity is generated by the digital nature on which (at least, at present) AI is built. *Nota bene*: there are, for now, no elements that "promise" the development of an AI species of analogical type, i.e., similar to natural intelligence;
- it is *non-judgmental*: the non-hermeneutic nature of AI makes it incapable of judgments (propositional assemblies of concepts) and even more of reasoning (propositional assemblies of judgments);
- is *non-comprehensive*: comprehensiveness refers to understanding, which is distinct (and, to a large extent, logically and psychologically opposed) to explanation – while explanation aims at causality and appeals to the intellect, comprehension aims at plausibility and appeals to intuition;
- is *non-creative*: AI can only make morphological combinatorial assemblies (permutations) based on known elements (*Nota bene*: here, the term knowledge has a purely mechanical meaning, not implying understanding);
- is of the type of *computation* (is computational): the operation of AI is purely based on computation – the primary logical model is that of the operation of the Turing machine, which is, essentially, an arithmetic calculation machine;
- it is purely *quantitative*: AI does not have access to qualities; although fundamentalist proponents of AI claim that increased computing power (and speed) will lead to the qualitative leap from non-consciousness to consciousness, in reality this leap is made by intuition, which is forbidden to AI;
- it is practically *unlimited operationally*: from a purely quantitative point of view, AI seems to be able to have unlimited growth in computing power, computing speed, memory capacity and the like;
- is *non-conscious*: AI will never have access to consciousness, i.e., to intuition and comprehension. The explanation of this impossibility resides in the digital character of AI, shown above – consciousness appears only under analogical conditions and through synergy, which necessarily attract comprehension (understanding) based on semantic ambiguity and cognitive uncertainty.

AI impact categories – theoretical issues

To assess the impact of AI, it must be analyzed the three categories of fundamental activities (*Nota bene*: typology suggested by Kantian philosophy): (a) *theoretical* activity – object-object relations, with both non-artifact objects, with the external observer, and without impact of observer on the object, namely, the so-called 1st-order cybernetic systems; (b) *praxeological* activity – subject-object relations, with the object not being an artefact, with the subject involved in the object, namely, the so-called cybernetic systems of the 2nd order; (c) *practical* activity – subject-subject relations, inter-acting within an artifact, namely, the so-called 3rd order cybernetic systems.

Awareness/consciousness

The question of the genesis of awareness from its biological basis is a problem that remains scientifically unsolved. Consequently, we can only discuss this matter, here, from a purely logical perspective. The theses we want to support are the following: (i) awareness/consciousness is not based quantitatively, but qualitatively; (ii) awareness/consciousness is related to analogical information/knowledge, not digital information/knowledge; (iii) awareness/consciousness is possible only through (or based on) representation.

(i) awareness/consciousness does not represent a qualitative leap as a result of quantitative accumulations – for example, by accumulating data/information/knowledge, above a certain threshold of this accumulation, consciousness arising is not necessary. If it were so, indeed, artificial intelligence could have made the leap to consciousness, because the quantitative accumulation, both as information memorization and as its processing (speed, degree of sophistication, etc.), is far superior to those found in human intellectual potential;

(ii) awareness/consciousness is possible exclusively from the perspective of analogical information – analogical information is information of a continuous type, relatively imprecise, involving overlaps or semantic mixtures and, especially, bearing connotations (in addition to definitional denotations) and, above all, involving ambiguity, the holistic view, and the intuition. As we have shown above, AI is possible (at least, for now) only on the basis of digital information, the latter having characteristics diametrically opposed to the analogical one;

(iii) the problem of representation is fundamental – representation is the ability of an entity to perceive another entity in the latter's physical absence. Perception conditioned by the actual presence of the perceived entity is a sensory (sensitive) perception, while perception not conditioned by the actual presence of the perceived entity is a representational perception. Representational perception (Dicker, 2011) is (presumed) specific to human being, so human being is credited with awareness/consciousness. *Nota bene*: the concept of awareness/consciousness is culturally loaded one. The substrate condition (in the case of human being, the psychological condition) for the emergence of consciousness is the awareness. The representational capacity of awareness makes it possible to integrate the cultural dimension into awareness, and this phenomenon of cultural integration (values, principles, "red lines", etc.) generates the very content of consciousness. The absence of awareness prevents, at the causal level, the generation of consciousness for AI.

So, from a logical perspective, awareness (and, *a fortiori*, consciousness) remains inaccessible to AI.

The unconscious/subconscious

From a logical point of view (*Nota bene*: a psychological point of view on AI is, of course, meaningless), in the absence of the conscious there is no unconscious (or subconscious). As is known, especially following behaviorist research and as a result of the development of neuroscience, the rationality of the human individual, especially the theoretical rationality, is based on the unconscious/subconscious. Psychoanalysis considers the unconscious either a "repository" of the individual's repulsions, or a conscious in a potential state (unactualized - *Nota bene*: here the polar terms potential - actual have their Aristotelian meaning. How the unconscious does not obey an operational rationality, so to speak (as is known, theoretical rationality refers to a rationality based on beliefs, propensities, preferences and the like, being a background rationality, not an inferential-analytical one, as is practical rationality, which is specific to the conscious), it follows that AI, respectively AID (artificial intelligence device) cannot benefit from the intuition proper to the unconscious and which has non-algorithmicity as its fundamental feature. On the contrary, AI is characterized by the exclusivity of algorithmicity.

The digital/analogical rapport

We have previously referred to the relationship between digital information and analogical information, as well as the differences and consequences of their processing. Here we will provide a more systematic set of distinctions, from a theoretical perspective:

(i) digital means binary („presence" vs. „absence", or "1" vs. "0") and it is imposed by the current level of technology based on the circulation of electric current in technical devices. Obviously, the digital is based on bivalent logic (the logic of the excluded middle). Although there are trivalent logics, and these logics can be combined with how the quantum states of matter ('0', '1' and 'x', where 'x' means indeterminate – , see Łukasiewicz's trivalent logic, and, for quantum indeterminacy, see Schrödinger's paradox), the use of these latter logics does not change things, because digital means discrete, non-continuous, and a logic, even if it is n-valent (with finite n), i.e., a logic of $(n+1)^{th}$ excluded, remains a discrete logic, perfectly similar to bivalent logic from a qualitative point of view, only quantitatively extended. *Nota bene*: even if n is infinite, if it is a natural number (or even a rational number), the logic in question remains of the discrete type, since both the set of natural numbers, and the set of rational numbers are countable sets, therefore discrete;

(ii) there is a proposal of non-discrete logic (fuzzy logic, of Zadeh), which is a logic of the power of the continuous (real number), but this logic is based on subjective evaluations of the numerical value of the membership function of the real set contained in the closed real interval [0,1]. Although fuzzy logic is massively used in computer programming, solving, in particular, many problems raised by uncertainty, it remains external to the substrate of calculation, so that this calculation remains discrete. Perhaps a *logic of similarity* (not membership, as is the case with fuzzy logic) might be more penetrating in this case and might ensure analogical (or, at least, quasi-analogical) AI behavior at the level of intrinsic computation, not only at the level of informational representation. There is also the possibility of developing a *logic of entanglement* (that is, of the non-separability of the local and the global between them) that could also approach analogical computing, through quantum computing;

(iii) analogical means the non-discrete, the continuous, the relatively imprecise. All these characteristics, which seem to reduce rigor, in fact, only reduce arithmomorphism (*Nota bene*: as this concept is introduced by the Romanian-American economist Nicholas Georgescu-Roegen), thus facilitating the genesis of novelty. Novelty results from semantic overlaps (either necessary or contingent), therefore, as said above, consciousness appears only as novelty, as synergy – so. novelty implies the analogical. The concept of novelty is a very demanding concept, so we will make some specific considerations in this regard:

- the novelty must present a difference in concept, not in degree, and even less in quantity;
- the difference in concept refers to the difference in meaning, i.e., referential (or denoted); – therefore, the novelty is revealed, first of all, from a semiotic perspective.

Nota bene: morphological combination, no matter how sophisticated (or even... ingenious), is not at all of the nature of novelty (AI can do morphological combinations to a much higher degree than natural intelligence, because this morphological combination is of the nature of calculation).

- semantic overlaps, "unnatural" associations, even cognitive or interpretation errors, are liable to produce novelty, because they leave the "jurisdiction" of the actual and enter that of the possible or, sometimes, of the (*prima facie*) impossible.

A question can be asked: does quantum computing change anything about digitality, which, as we know, is incompatible with the existence of consciousness? We make a few comments on the matter:

- quantum computing is based on the quantum property called linear superposition, i.e., the existence of uncertainty about a result before it is measured (the famous example is that of Schrödinger's cat);
- the idea is that, before signifying either "0" or "1", any state signifies either "neither 0 nor 1" or "both 0 and 1", and only the actual measurement (or, according to some researchers, the appearance, in a human consciousness, of even just the intention to measure) causes the state of uncertainty to collapse into either the "0" state or the "1" state, but the final outcome of the collapse cannot be predicted;
- it is easy to see that in the state, let us say, "indeterminate (*Nota bene*: this is the third truth value in Łukasiewicz's trivalent and ... discrete logic) we do not have an infinity of values (as in the case of fuzzy logic) but only an indeterminacy (temporal or conditional), so that, when verifying precisely determined

conditionalities, the indeterminacy collapses, as said, into one of the two...discrete values;

- therefore, from the point of view of the interest of analysis in the present study, quantum computing does not, in any way, transform digitality into analogicality – the effect is related, above all, to the computing power, i.e., to a quantitative aspect, not to a qualitative one;
- in conclusion, quantum computing, i.e., QAID – quantum artificial intelligence devices (Wichert, 2024) does not make the transition to analogical computing, so it does not bring any leap from the perspective of the problem of AI consciousness.

Genetic code/machine code rapport

Most AI fundamentalists accept without too much grounding, an equivalence between the natural genetic code, which programs the functionality and behavior of the living biological entity, and the programs (either hardware or software) that ensure the functionality and behavior of AI machines (i.e., logically living systems). Of course, this equation is not... "innocent": it tries to inculcate the idea that the human person and the AI are species of the same genus, with the only (presumably non-essential) difference that the human person's "hard" is biological in nature, and the AI hard is technological in nature, thus resulting in the "soft" of both is equivalent, at least functionally. In our opinion, there are, however, fundamental differences between human being and AI from a "soft" perspective even if we accept, for the sake of discussion, the secondary character of hard (*Nota bene*: the secondary character of hard can be accepted if we "make sure" that the hard structure of natural life and the hard structure of artificial life are isomorphic between them. As structure generates function, it follows that structural isomorphism is liable to generate functional isomorphism).

Before listing some of the differences that we consider crucial between genetic code and machine code, we briefly try to clarify the two concepts:

(i) *genetic code* – although "recent" biologists no longer accept the idea that by genetic code we must understand an algorithmic program of behavior, i.e., of generating functions, as actions or reactions, a determining role is still accepted, causally and/or conditionally, for the genetic code in the behavior of biological life. So, the genetic code can be understood as a biological hard on the basis of and through which the software (for example, education) generates and maintains functions: decision/choice, action, evaluation, etc.;

(ii) *machine code* – represents the program, "written" in the language accessible to the machine (*Nota bene*: in no case should it be said "in the language understood by the machine") which, analogously to the case of biological life, based on and through the hard mechanical (the term mechanical refers to anything non-biological, even if, for example, hard is purely an electronic technology) generates and maintains machine functions of any kind.

The critical differences between genetic code and machine code are as follows:

- the *non-deliberative* character of the mutations produced in the genetic code (*Nota bene*: recently, based on genetic engineering, we can speak of a mixed or hybrid, random-deliberative character), compared to the *deliberative* character of the "mutations" produced in the machine code ;
- the *random* character (not stochastic!), of the mutations involved in the genetic code, while the machine code undergoes *deterministic* mutations;
- in the case of genetic code, we have no predictability of mutations, while, in the case of machine code, we can have such predictability. We develop this assertion more analytically: predictability has two meanings here: (a) predictability of mutation production; (b) predictability of mutation occurring:
  (a) <u>predictability of mutation production</u>: in the case of the genetic code there is no such predictability (except interventions of genetic engineering which, at least for the moment, are insignificant in weight and frequency); also, in the case of machine code, we do not have this type of predictability, because no

one can predict when a (deliberative) mutation will occur in the machine code;

(b) predictability of mutation that occurs: in the case of the genetic code, the unpredictability of the mutation (type, extent, timing, etc.) is maintained; instead, in the case of the machine code, we can have predictability: when a mutation is designed, the type, extent, impact etc. of it are well-known – they can even be part of a program or a strategy, which aims decrease the unpredictability;

- mutation in the genetic code is non-artefactual (or, through genetic engineering, partially artefactual), while mutation in the machine code is exclusively artefactual.

Free will

Free will is a species of the genus we call freedom. It is defined as the *unpredictable freedom of opposition to non-natural necessity*, most often to social necessity, but it can also refer to individual cultural necessity. It is obvious that there can be no opposition to natural necessity. Opposition to social necessity (for example, to social norms) is possible exclusively for living entities endowed with consciousness, that is, as previously shown, exclusively for human being (natural intelligence). A few analytical considerations are useful at this point:

- one might object to the "verdict" that only natural intelligence is capable of free will on the following reasoning: to oppose a non-natural necessity is, in advance, to make the decision about this opposition, that is, to decide to oppose. Decision means choice (otherwise, it is natural necessity) and choice is a result/effect of intelligence, so the choice could also be the result of AI which is obviously a species of intelligence;
- we object to the previous reasoning as follows: AI can only "reason" on the basis of the deterministic program that gives it its very existence. Now, this program is perfectly predictable – if from it a conforming behavior to the norms is inferred, then we have no opposition and, *a fortiori*, we have no manifestation of free will, but if from it an elusive behavior is inferred, then we have opposition, it is true, but perfectly predictable, so again we have no free will;
- however, one could object directly against the premise in mind, namely against the qualification of free will as a freedom of *unpredictable* opposition to non-natural necessity. We defend this qualification as follows:
  - the non-natural necessity is, in turn, an artefact, so nothing would prevent one who designed the non-natural necessity from considering, in this design, the possibility of the human individual's opposition to the functioning of that artefact, and thus annihilate it *ante factum*;
  - the fact that this *ante factum* annihilation did not occur means that free will was not (in principle, is not, in fact) predictable. We therefore conclude that AI cannot have free will, since any behavior (choice, action, opposition, etc.) is programmed, i.e., both compliance and opposition to non-natural necessity are (perfectly) predictable, as they are inferable logical.

Self-learning

The above considerations on free will could be overturned by the phenomenon that has been termed *self-learning*. Self-learning is constitutive of the human being (and for that matter of any living entity in the biological sense), so this concept will be discussed here only from an AI perspective. Self-learning generally refers to the ability of an AI-powered machine to update program (the software) it operates, as a result of "inferences" drawn by the AI itself (not by its human creator or the individual human monitoring it), from the very running of the program in question. *Nota bene*: updating means the three operations: (a) deleting instructions; (b) adding instructions; (c) change of instructions. We will discuss two important issues in this issue: (i) the

possibility of self-learning;
(ii) self-learning modality:

(i) AI is a technological device (a biological device version of AI is, at least for now, utopian if not... dystopian) so it operates by necessarily following the instructions of its program (its machine code). So, the AI device could, of course, learn from its own operation and, as a result, could update (in the above sense of the term update) the own program if it... already has instructions to do so. In other words, it is not about self-learning at all, but about an update of the program also done by the human operator, so, only this update is introduced, in an anticipatory way, into the program. In fact, it is only an automation, let's call it, of the 2nd order, that is, an automation of automation, but everything is foreseen, designed and implemented by human being – it is, to use more "civilian" words, a convenience of the human user who, instead of drawing conclusions from the operation of the AI and then updating them in the program, transfers this task to the AI device, but the latter proceeds perfectly deterministically and within the framework of the necessity with the update in question. Bottom line, AI never learns anything, it is taught by humans. There is, here, a confusion analogous to the payment of VAT in the economy: VAT is collected and paid to the public budget by the seller, but it is included in the price and, therefore, borne by the consumer (more precisely, the final consumer). We cannot say that the VAT is paid by the seller: in fact, the consumer has "authorized" the seller, through the fiscal procedure norm, to transfer the VAT amount to the state budget. Likewise, the human "authorizes" the AI device to operate the appropriate update in the program, and this authorization is already contained in the program. Without self-updating instructions, the so-called self-learning would obviously be impossible

(ii) the real issue here is another one: could the AI device (AID) operate updates in its own program without the program in question containing instructions for that update? In other words: could AID, at some point in its technological development, ignore or even oppose its own program so that it could update it without instructional permission? Current fears, voiced by more or less AI "knowers" (who anticipate apocalypses such as AI turning against humans and other similar phenomena) favor an affirmative answer to this question. Of course, this affirmative answer is erroneous for the following reasons:
- to make its own updates in the program (i.e., excluding the pre-programmed ones) the AID must exhibit free will, because it must oppose the non-natural necessity that is contained in its operating program. But AI is incompatible with free will, as previously shown, so self-learning, as self-learning alongside or against the instructions of own program is impossible in principle, not accidentally;
- only one possibility remains (which has a non-zero probability), namely the error. We can discuss the possibility of this error from two perspectives:
  - (type α error) error in the AID program that allows AID to operate updates alongside or against the program. It is obvious that, in this case, the AID does not operate on the basis of free will, but also on the basis of the program. Although, in principle, a self-learning process takes place, in fact, this process is still based on the program, more precisely, on errors in the program, so we return to the previous case – it can be said that self-learning is programmed by error, i.e., it is what we called error of type α;
  - (type β error) there is no error in the program (so there is no possibility of self-learning) but, during the execution of that program, errors occur in the operation of the program in question, errors that may result in the activation of the possibility of AID to operate, "on its own" updates in the own program. Logically, although less obviously than in the previous case, we are in the same situation: the malfunctioning of the program is due to a poor quality of the program in question, namely, the fact that its instructions contain this possibility: of malfunctioning. Consequently, even in this case it cannot be about self-learning in the strong sense of the

term, but, as in all previous cases, in a weak or narrow sense: self-learning is pre-programmed, in a one way or another.

A final issue that can be raised here is this: couldn't all these programming errors lead to the situation where, upon sensing them, the AID switches to genuine self-learning? For example, it could correct these errors and thus block... self-learning. Or, conversely, they could operate updates that remove any programmatic obstacles to self-learning, thus initiating a vicious (or virtuous, as the case may be) cycle of self-learning. The answer is obvious and close at hand: to do this, AID must be capable of understanding (comprehension), which, as we have shown before, is a principled no-go "territory" for AI.

Protection of the human being

The question of protecting the human being (and humanity in general) in the face of possible AI empowerment, considered globally, which empowerment could challenge the human right to control AID and, symmetrically, claim the AI "right" (specifically, of AID) to control human being (and humanity as a whole) is a consequence of self-learning. However, the matter raises some specific questions, therefore we will make the following analytical additional considerations:

- the protection of the generic human being (individual, respectively mankind) cannot be ensured (guaranteed, respectively defended) by the standard (orthodox) normative framework, because AIDs do not obey this normative framework – for example, a formulated normative clause has no effectiveness as follows: "*it is forbidden, for any AID, to physically attack any human individual, in any circumstance, in any way and for any reason*", possibly the clause could be accompanied by a sanction (*Nota bene*: the issue of the type of sanction, of its gradation proportional to the "fact" etc., to be applied/applicable to an AID is an issue in itself that would deserve to be dealt with separately – we will not develop this discussion here);
- it remains that the protection of the generic human being is ensured (guaranteed, respectively defended) by means of the program (machine code) that organizes and directs the operation of an AID. The program instructions related to the protection of the generic human being are contextual instructions, i.e., contextualizing the operation/behavior of the AID, a kind of *conditio sine qua non* master instructions. For example, they should, in the first instance, select those actions (either acts or abstentions) which have as their existential constraint, so to speak, that of protecting, in the broadest sense of the term, the generic human being, then, in a second instance, it should be able to block an "illegal" AID action, and, in a third instance, to disable the AID, and then, in a fourth instance, proceed to the self-destruction ("suicide") of the AID in question. We believe that such hierarchical protection on four levels is of a nature, in principle, to ensure the protection of the generic human being in his/her relations with AID.

Nota bene: of course, analogously to the case of errors discussed above (self-learning), here too errors can occur in the operation of the program, namely the same types of errors: (a) of type α: erroneous instructions that does not protect the generic human being in the face of actions (acts or abstentions, as the case may be); (b) of type β: instructions which are correct from the point of view of protecting the generic human being but which, by chance, function erroneously from that point of view. Of course, in the case of generic human protection, a third type of error (let's call it γ-type error) can occur which refers to instructions that either permit or directly indicate the physical aggression of the AID on the human being (for example, if AID were to be used in wars, such instructions become ... mandatory, of course accompanied by criteria for recognizing enemies, so that aggression is not exercised on combatants from the same „tribe" – but, obviously, within the error of type γ errors of types α or β may occur).

AID protection

Is there a problem called AID protection? We believe that the answer to this question is affirmative, and we wish to address the following issues: (a) the protection of AID in relation to people; (b) protection of AID in relation to other AID.

- the protection of AID in relation to people
  - refers to cases where an AID is subject to human manipulation of false behaviors and actions (e.g., simulating a meeting where a decision is made that is advantageous to the manipulators). In this case, protection can be ensured at an ethical level, regarding human individuals (especially through education) but also the normative framework can be useful, because people are subject, through the Constitution and the laws derived from it, to sanctions of various categories, intensities and purposes, applied by society;
  - protection of AID in relation to other AID – refers to cases where some AIDs can exert distorting effects, from the role, operation and programmed behavior, on other AIDs. These effects can be accidental but they can also be systematic. There are, of course, two distinct cases:
    - the mentioned effects are implicit in the instructions of the programs, so they can be corrected by updating those programs;
    - the mentioned effects represent "products" of self-learning. In this case, as it follows from the above, the so-called self-learning is based, in one way or another, on the instructions.

Conclusions

It is worth discussing, here, the problem of the AID "coalition" for the establishment of domination over people, evoked previously. It is obvious that such an eventuality involves the problem of the protection of AID in relation to other AID, because there is an autonomization of direct communication between AIDs, separate from and outside the observability and controllability of humans. Here, too, there is no other solution than that of the impact analysis of the programs with which AID is equipped. In our opinion, a theory and a methodology that could be called AI hermeneutics will have to be developed (and will be developed, under the pressure of facts), which will be a hermeneutic of a logical, hypothetic-deductive type, obviously much simpler than hermeneutic of philosophy with which people operate. We believe that it is theoretically possible to trigger an "arms race" between the possibility of autonomous intra-AI communication in relation to human – AI communication. *Nota bene*: it must not be forgotten that, unlike the case of philosophical hermeneutics, logical hermeneutics or deductive hermeneutics is perfectly within the "reach" of AI.

References

Dicker, G. (2011). The representational theory of perception and the problem of perception. In Oxford Scholarship online, https://doi.org/10.1093/acprof:oso/9780195381467.003.0003.

Dinga, E. (2020). Logically living system: A generative machine for auto-poietic systems. In M. Pańkowska (ed.) *Handbook of Research on Autopoiesis and Self-Sustaining Processes for Organizational Success*. IGI Global Publishing Tomorrow's Research Today, ISBN: 9781799867135. https://www.igi-global.com/book/handbook-research-autopoiesis-self-sustaining/256641

Dinga, E. (2024). *Economic Resilience During Overlapped Crises – Antifragility, Sustainability and Autopoieticity*, Palgrave Macmillan – Springer Nature. In processing.

Kreuzer, T., Papapetrou, P., Zdravkovic, J. (2024) – Artificial intelligence in digital twins. A systematic literature review, *Data & Knowledge Engineering*, vol. 151, https://doi.org/10.1016/j.datak.2024.102304.

Wichert, A, 2024 – *Quantum Artificial Intelligence with Qiskit*, Chapman and Hall/CRC, ISBN: 9781003374404. https://www.routledge.com/Quantum-Artificial-Intelligence-with-Qiskit/Wichert/p/book/9781032448978?srsltid=AfmBOooKL2JhzI3VDwR1wCktNKt4gR5TcURt7BHpy-CEcqcP_jknYtHs